Bootstrap-Based Bias Correction for Graphical Estimators from Probability Plots with Complete and Right-Censored Data

Somboonsavatdee, Anupap
Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University
Phayathai Road
Bangkok 10330, THAILAND
anupap@acc.chula.ac.th

Introduction

Being a popular graphical tool for assessing parametric distributional assumption, for data from (log)location-scale, the probability plot can also be used to estimate the location and scale parameters by fitting a line through the plot (see Section 6 in Meeker & Escobar 1998). The parameters can be easily estimated through this graphical method especially for censored data. The property of the graphical estimators (GEs) and their comparison to maximum likelihood estimators (MLEs) are discussed in many censoring configurations (Nair 1984, Somboonsavatdee et al. 2007, Genschel & Meeker 2010, and Olteanu & Freeman 2010). Besides being theoretically less efficient to MLEs in most cases (Nair 1984, Somboonsvatdee et al. 2007, and Genschel & Meeker 2010), the GEs are still popular used among reliability engineers.

Just like any other estimators, the GEs can be biased, therefore, in this study, we have look into how we would correct the biases of the GEs through bootstrap method. There have been some previous works on the reduction of bias for the MLEs (Thoman & Bain 1984, and Hirosi 1999) and GEs (Zhang et al. 2006) but, not through bootstrap method. Due to the limitation of number of pages allowed in this paper, we mainly focus on the results for smallest extreme value (SEV) distribution (or Weibull for log-location-scale distribution) with complete data and Type-II censored (also known as failure-censored) data where the results for (log)normal distribution are briefly discussed in the discussion section.

Graphical Estimators for (Log-)Location-Scale Distribution

The commonly known location-scale distributions are normal, logistic, and SEV distributions (or lognormal, loglogistics, and Weibull distributions for log-location-scale distributions). If $F(\mu, \sigma)$ is a location-scale distribution with μ and σ being its location and scale parameters, it follows that $x_q = \mu + \sigma F_{0,1}^{-1}(q)$ where x_q is the $q-$th quantile from $F(\mu, \sigma)$ and $F_{0,1}^{-1}(\cdot)$ is the quantile function of $F(0, 1)$. With this property, the ordinary least squares (OLS) estimators for intercept and slope from the probability plot, by having sample order statistics on y-axis and their corresponding theoretical quantiles from $F(0, 1)$ on x-axis, can be used as the estimators of the location and scale parameters.

As the focus of this study will be on the SEV distribution with complete and Type-II censored data, the GEs is simply the OLS estimators from a linear plot of $\{F_i,\ X_{(i)}\}$'s for $i = 1, 2, \ldots, r \leq n$ with $X_{(i)}$ is the $i-$th sample ordered statistics and $F_i = -\log(-\log(q_i))$ where $q_i = \frac{i-.5}{n}$ is a mean-rank plotting position, r is the number of uncensored data, and n is sample size. Therefore, the GEs are

$$\hat{\sigma} = \frac{\sum_{i=1}^{r}(X_{(i)} - \bar{X}_r)(F_i - \bar{F}_r)}{\sum_{i=1}^{r}(F_i - \bar{F}_r)^2} \quad \text{and} \quad \hat{\mu} = \bar{X}_r - \hat{\sigma}\bar{F}_r$$

where $\bar{X}_r = \frac{1}{r}\sum_{i=1}^{r} X_{(i)}$ and $\bar{F}_r = \frac{1}{r}\sum_{i=1}^{r} F_i$. One can also use different choice of the plotting position such as median-rank plotting position (Barnett 1975, and see page 276 in Nair & Somboonsvatdee 2010) but, the results should not be significantly different especially for large sample.

Bootstrap-Based Bias Correction

The bias of the estimator $\hat{\theta}$ of a parameter θ can be computed by $Bias_{\theta}(\hat{\theta}) = \mathbf{E}(\hat{\theta}) - \theta$. As we can see that the bias of an estimators can only be computed when the true value of the parameter is known, however, the bias can also be estimated by using the popular bootstrap method (Efron & Tibshirani 1994). The bootstrap-based estimated bias can be computed as below,

$$\widehat{Bias_{\theta}}(\hat{\theta}) = \left(\frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^B\right) - \hat{\theta}$$

where $\hat{\theta}_j^B$'s are generated through the bootstrap simulation. Then, the estimator are bias-corrected by subtracting the estimated bias from the estimator,

$$\tilde{\theta}_{bbc} = \hat{\theta} - \widehat{Bias_{\theta}}(\hat{\theta}) = 2\hat{\theta} - \left(\frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^B\right)$$

where $\tilde{\theta}_{bbc}$ is the bootstrap-based bias corrected estimator for θ.

In this study, it is only appropriate to use parametric bootstrap method since we are dealing with (log-)location-scale distribution which is parametric distribution. The parameters of interested are location (μ) and log-scale [$\log(\sigma)$] parameters so that performances of the estimators are independence the value of μ and $\log(\sigma)$, so we need only consider the case $\mu = 0$ and $\sigma = 1$. The (bootstrap-based) bias corrected graphical estimators (BCGEs) are computed as described above.

Simulation Results

The results are based on the simulation with the simulation size N of 5000. The data are randomly generated from standard SEV distribution ($\mu = 0$ and $\sigma = 1$) with (Type-II) censoring proportion $p = 0$ (no censoring), 0.2, .04, 0.6, 0.8 and sample size n =10, 25, 50, 75, 100 . For each combination of censoring proportion p and sample size n, 5000 GEs of μ and $\log(\sigma)$ are obtained, then for each pair of GEs of μ and $\log(\sigma)$, the parametric bootstrap with bootstrap simulation size B of 5000 is used in order to estimate their biases, then used to obtain the corresponding BCGEs of that pair of the GEs.

Tables 1 and 2 show the simulation results for GEs and BCGEs of location (μ) and log-scale [$\log(\sigma)$] parameters. The results includes bias, variance, mean square error (MSE) of the estimators. Moreover, the results in the tables also include the two versions of relative efficiency (RE) for comparing the efficiency of the BCGE to the GE. RE1 is computed by the ratio of variances of GE to BCGE and RE2 is similarly computed by the ratio of MSEs. If the relative efficiency is greater than 1, this implies that that BCGEs are more efficient than GEs; if less than 1, this implies that BCGEs are less efficient; and if equal to 1, this implies that both estimators are equally efficient.

Let consider the simulation results for the estimators of location parameter (μ) from Table 1. We can see that the biases of BCGEs are very close to 0 in all scenarios. This is to be expected

Table 1: Estimation of Location Parameter μ — Type-II Censoring

		$(\hat{\mu}) \rightarrow$ GE			$(\hat{\mu}) \rightarrow$ BCGE				
Censoring Proportion p	Sample Size n	Bias	Var	MSE	Bias	Var	MSE	RE1	RE2
0	10	-0.054	0.119	0.122	-0.002	0.118	0.118	1.01	1.04
	25	-0.029	0.048	0.049	-0.001	0.048	0.048	1.01	1.03
	50	-0.014	0.024	0.024	0.000	0.024	0.024	1.00	1.01
	75	-0.010	0.016	0.016	0.000	0.016	0.016	1.00	1.01
	100	-0.003	0.012	0.012	0.000	0.011	0.011	1.00	1.00
0.2	10	-0.055	0.163	0.166	-0.002	0.167	0.167	0.98	0.99
	25	-0.019	0.066	0.066	0.000	0.066	0.066	0.99	0.99
	50	-0.010	0.033	0.033	0.000	0.033	0.033	0.99	1.00
	75	-0.009	0.023	0.023	0.000	0.023	0.023	0.99	1.00
	100	-0.007	0.017	0.017	0.000	0.018	0.018	1.00	1.00
0.4	10	-0.058	0.294	0.298	-0.003	0.311	0.310	0.95	0.96
	25	-0.032	0.128	0.129	-0.001	0.132	0.132	0.97	0.98
	50	-0.023	0.067	0.067	0.000	0.068	0.068	0.98	0.99
	75	-0.019	0.046	0.046	0.000	0.047	0.047	0.98	0.99
	100	-0.008	0.036	0.036	0.000	0.036	0.036	0.99	0.99
0.6	10	-0.100	0.719	0.729	-0.007	0.794	0.794	0.91	0.92
	25	-0.039	0.310	0.311	-0.001	0.322	0.322	0.96	0.97
	50	-0.031	0.172	0.173	0.000	0.178	0.178	0.97	0.97
	75	-0.029	0.112	0.113	0.000	0.116	0.116	0.97	0.98
	100	-0.028	0.087	0.088	0.000	0.090	0.090	0.97	0.98
0.8	10	-0.163	3.302	3.328	-0.014	3.795	3.794	0.87	0.88
	25	-0.116	1.233	1.246	-0.007	1.360	1.359	0.91	0.92
	50	-0.073	0.684	0.689	-0.002	0.727	0.726	0.94	0.95
	75	-0.075	0.477	0.482	-0.002	0.507	0.507	0.94	0.95
	100	-0.046	0.390	0.392	-0.001	0.405	0.405	0.96	0.97

since BCGEs are the bias corrected estimators from GEs. In most scenarios, except when $p = 0$, the variance of BCGEs are larger than those of GEs. It is expected that the variances of BCGEs would be larger than those of GEs due to the effect from the estimated biases. Interestingly, at censoring proportion $p = 0$, the variances of BCGEs are less than or equal to those of GEs. In scenarios with some censoring ($p \neq 0$), the MSEs of BCGEs are larger than those of GEs, however, the MSEs of BCGEs are getting closer or equal to those of GEs when the sample size n is getting larger. With no censoring (p=0), RE1 and RE2 are greater than 1 for smaller samples ($n \leq 25$ for RE1 and $n \leq 75$ for RE2) but only by little, and are equal to 1 for larger sample ($n \geq 50$ for RE1 and $n = 100$ for RE2). With some censoring ($p \neq 0$), RE1 and RE2 are less than 1 especially from small sample, then getting closer to 1 as the sample size increases. Moreover, at fixed sample size n, the RE1 and RE2 are decreasing as the censoring proportion increases. By considering values of RE1 and RE2, the BCGEs are somewhat equally efficient or a bit more effcient to GEs when used for complete data, however, the BCGEs are less efficient than GEs when the censoring proportion p increases and when the sample size n is small.

Table 2 shows the simulation results for the estimators of log-scale parameter $[\log(\sigma)]$. We can obviously see that the biases of BCGEs are all zeroes. This is due to the fact that the estimated bias is independent to both values of location (μ) and scale parameter (σ). The variances for BCGEs and GEs are the same which leads to the values of MSEs of BCGEs being less than those of GEs. With equal values of variances, the RE1 are equal to 1 for all scenarios, and with smaller values of MSEs of BCGEs, the RE2 are all greater than 1. Overall BCGEs performs better than GEs either using the criteria of variances or MSEs due to the fact that the BCGEs are unbiased but the variances are the same as those of GEs.

Table 2: Estimation of Log-Scale Parameter* $\log(\sigma)$ — *Type-II Censoring

		$[\log(\hat{\sigma})] \rightarrow$ GE			$[\log(\hat{\sigma})] \rightarrow$ BCGE				
Censoring Proportion p	Sample Size n	Bias	Var	MSE	Bias	Var	MSE	RE1	RE2
0	10	-0.082	0.103	0.110	0.000	0.103	0.103	1.00	1.07
	25	-0.040	0.040	0.042	0.000	0.040	0.040	1.00	1.04
	50	-0.022	0.020	0.021	0.000	0.020	0.020	1.00	1.02
	75	-0.015	0.014	0.014	0.000	0.014	0.014	1.00	1.02
	100	-0.013	0.010	0.011	0.000	0.010	0.010	1.00	1.02
0.2	10	-0.113	0.150	0.163	0.000	0.150	0.150	1.00	1.08
	25	-0.051	0.064	0.067	0.000	0.064	0.064	1.00	1.04
	50	-0.024	0.032	0.032	0.000	0.032	0.032	1.00	1.02
	75	-0.021	0.022	0.022	0.000	0.022	0.022	1.00	1.02
	100	-0.016	0.017	0.017	0.000	0.017	0.017	1.00	1.01
0.4	10	-0.157	0.234	0.258	0.000	0.234	0.234	1.00	1.11
	25	-0.069	0.094	0.098	0.000	0.094	0.094	1.00	1.05
	50	-0.041	0.049	0.050	0.000	0.049	0.049	1.00	1.03
	75	-0.029	0.033	0.034	0.000	0.033	0.033	1.00	1.03
	100	-0.020	0.025	0.025	0.000	0.025	0.025	1.00	1.02
0.6	10	-0.254	0.410	0.475	0.000	0.410	0.410	1.00	1.16
	25	-0.099	0.151	0.161	0.000	0.151	0.151	1.00	1.06
	50	-0.053	0.077	0.080	0.000	0.077	0.077	1.00	1.04
	75	-0.043	0.051	0.053	0.000	0.051	0.051	1.00	1.04
	100	-0.036	0.040	0.041	0.000	0.040	0.040	1.00	1.03
0.8	10	-0.656	1.601	2.031	0.000	1.601	1.600	1.00	1.27
	25	-0.225	0.356	0.406	0.000	0.356	0.356	1.00	1.14
	50	-0.114	0.166	0.179	0.000	0.166	0.166	1.00	1.08
	75	-0.086	0.109	0.116	0.000	0.109	0.109	1.00	1.07
	100	-0.065	0.086	0.091	0.000	0.086	0.086	1.00	1.05

Discussion

From the simulation results, we can see that the bootstrap-based bias correction improve the performance of the graphical estimator for log-scale parameter $[\log(\sigma)]$, however, not for location parameter. All of the results are due to the effects from the bootstrap-based estimated bias which are discussed below.

Estimated Bias for $\log(\hat{\sigma})$

Suppose $X_1, X_2, \ldots, X_n$ are *iid* random variables from $SEV(\mu, \sigma)$. Let $\sigma_1 = 1$, with $\hat{\sigma}$ described in earlier section, it follows that $\hat{\sigma} = \sigma \times \hat{\sigma_1}$ and $\log(\hat{\sigma}) = \log(\sigma) + \log(\hat{\sigma_1})$ where $\hat{\sigma}_1$ is computed from $X_1^*, X_2^*, \ldots, X_n^*$ with $X_i^* = \frac{X_i - \mu}{\sigma}$. Therefore, it follows that

$$Bias_{\log(\sigma)}[\log(\hat{\sigma})] = \mathbf{E}\{\log(\sigma) + \log(\hat{\sigma}_1)\} - \log(\sigma) = \mathbf{E}\{\log(\hat{\sigma}_1)\} = Bias_{\log(\sigma_1)}[\log(\hat{\sigma}_1)].$$

We can see that the $Bias_{\log(\sigma)}[\log(\hat{\sigma})]$ is independent to both values of μ and σ, and can be accurately estimated through the bootstrap with large number of bootstrap simulation. With the sample size n and censoring proportion p being consistent with the simulations in this study, one can use the values from Table 2 to get the bias of $\log(\hat{\sigma})$ and to compute BCGE of $\log(\sigma)$. For example, at sample size of $n = 50$ with censoring proportion $p = 0.4$, if the GE for $\log(\sigma)$ is 1.6, then from Table 2, its corresponding bias is -0.041. Therefore, the corresponding BCGE for $\log(\sigma)$ is 1.6 - (-0.041)= 1.641.

Estimated Bias for $\hat{\mu}$

Similarly, suppose $X_1, X_2, \ldots, X_n$ are *iid* random variables from $SEV(\mu, \sigma)$. Let $\mu_0 = 0$, with $\hat{\mu}$ described in earlier section, it can be shown that $\hat{\mu} = \mu + \hat{\sigma}\hat{\mu_0}$ with $\hat{\mu}_0$ is computed from $X_1^*, X_2^*, \ldots, X_n^*$.

Therefore, it follows that

$$Bias_{\mu}(\hat{\mu}) = \mathbf{E}\{\mu - \hat{\sigma}\hat{\mu}_0\} - \mu = \mathbf{E}\{\hat{\sigma}\hat{\mu}_0\}.$$

We can see that the bias depends on the value of $\hat{\sigma}$ which is a function of the data and, more importantly, which is random. Therefore, while the bias of the estimator is being adjusted, the variance of the estimator will increase as seen in the results in Table 1. Unlike the bias for $\log(\hat{\sigma})$, the bias for $\hat{\mu}$ cannot be accurately estimated through the bootstrap simulation. This is due the fact that the value of $\hat{\mu}$ and $\hat{\sigma}$ from the sample will be treated as the parameters for generating the bootstrap simulation, and of course, the value of $\log(\hat{\sigma})$ from the sample will definitely affect how well the estimation of the bias of $\hat{\mu}$ is.

Other factors that influence the performance of the BCGEs are the sample size n and the censoring proportion p. Since at the sample sample size n and large censoring proportion p, the GEs tend to have large absolute value of the biases, therefore, by bootstrap-based bias corrected, this would improve the performance of the graphical estimators especially for log-scale parameter $[\log(\sigma)]$.

As mentioned earlier that due to the limitation of the number of pages, we only focus the results from SEV distribution. We have also done some simulation study for (log)normal distribution with the same set of sample size n and censoring proportion p. The results are very similar to those of SEV distribution where BCGEs do better than GEs for log-scale parameter $[\log(\sigma)]$ estimation but, do worse than GEs or do about equally well for location parameter μ) estimation. The results should be similar for all (log-)location-scale distributions for the data under Type-II censoring.

We want to emphasize that the intention of this study is not to promote the use of the graphical estimators over the MLEs or other types of estimators, but to explore how would we further improve the performance of the graphical estimation method, and to study its property. To conclude this study, with complete or type-II censored data, if one choose to use the graphical estimators, our suggestion is to use BCGEs over GEs only for estimating (log-)scale parameter $[\log(\sigma)]$ especially at small sample size n and high censoring proportion p.

REFERENCES

Barnett, V. (1975). Probability plotting methods and order statistics. *Journal of the Royal Statistical Scociety – Series C: Applied Ststistics,*, 24(1):95-108.

Escobar, L. (2010). Comments. *Quality Engineering*, 22:284-288.

Efron, B., and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap. Boca Raton*, FL: CRC Press.

Genschel, U. and Meeker, W. Q. (2010). A comparison of maximum likelihood and median-rank regression for Weibull estimation. *Quality Engineering*, 22:236-255.

Hirosi, H. (1999). Bias correction for maximum likelihood estimates in the two parameter Weibull distribution. *IEEE Transactions on Dielectrics and Electrical Insullaion*, 6:66-68.

Nair, V. N. (1984). On the behavior of some estimators from probability plots. *Journal of the American Statistical Association*, 79:823-831.

Nair, V. N., and Somboonvatdee, A. (2010). Comments. *Quality Engineering*, 22:273-277.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data.* New York: John Wiley & Sons.

Olteanu, D. and Freeman, L. (2010). The evaluation of median-rank regression and maximum likelihood estimation techniques for a two-parameter Weibull distribution. *Quality Engineering*, 22:256-2010.

Somboonsavatdee, A., Nair, V. N., and Sen, A. (2007). Graphical estimators from probability plots with right-censored data. *Technometrics*, 49:420-429.

Thoman, D. R., and Bain, L. J. (1984). Inferences on the parameters of the Weibull distribution. *Technometrics*, 11:445-460.

Zhang, L. F., Xie, M. and Tang, L.C. (2006) Bias correction for the least squares estimators of Weibull shape parameter with complete and censored data. *Reliability Engineering and System Safety*, 91:930-939.

RÉSUMÉ

Anupap Somboonsavatdee is Lecturer at the Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, Thailand. He received his PhD in Statistics from the University of Michigan.

ABSTRACT

Probability plots are popular graphical tools used by reliability engineers and other practitioners for assessing parametric distributional assumptions. When used for data from (log)location-scale families, the location and scale parameters can be estimated by fitting a line through the plot. This method is quick-and-easy especially when used for censored data. The commonly known problem of this graphical estimation is its bias. In this study, we approximate the bias through the bootstrap method, and then make adjustment to the graphical estimators. The properties of the bootstrap-based bias corrected graphical estimators are studied through simulation for selected distributions with complete and data and right-censored data under Type-II censoring. We find that with bootstrap-based bias correction, the efficiency of the graphical estimators is improved for log-scale estimation especially for small sample size and high censoring proportion. The effect from the censoring proportion on the improvement of the estimators and the relationship between true bias from a sample and its approximated bias are also studied and discussed.